Web-at-Risk: A Distributed Approach to Preserving our Nation's Political Cultural Heritage

Overview

The California Digital Library has been awarded a three-year, $2.4 million grant from the Library of Congress as part of the National Digital Information and Infrastructure Program (NDIIPP) to develop a web archiving service that will be used by libraries to capture, curate, and preserve collections of web-based government and political information. This literature is a critical element of our nation's heritage and is increasingly found exclusively online, putting it at greater risk of being lost. The collections will focus on local political activities and movements, such as the California gubernatorial recall election of 2003.

Specifically the core functions of the web archiving service will enable curators to build, manage and make accessible through search or browse collections of web-based materials. This includes providing the ability to perform the following activities related to archiving web based content: collect, monitor, QA, manage, describe, search, browse, display, store, preserve and manage rights.

The development of a web-archiving service is a practical undertaking. This is a "proof-of-concept" project. The tools built for the web-archiving service will be open source, extensible, and modular. This is essential in that the components must be configurable, modular, and extensible enough so as to be exploited by the workflows and repositories of our Web-at-Risk project partners. In addition, the service must be complimentary with international web-archiving efforts currently underway. Finally, the web archiving service will be designed to support and augment the workflows currently used by curators who are actively building collections.

Project Partners

The project involves a matrix of project partners with CDL as the lead partner. ***Main partners*** are funded and will assist in the establishment of the web archiving service and will take responsibility for specific aspects of the service. As such, staff working at partner institutions will work with CDL staff to implement the service. In addition, collected and curated content will be housed at their respective institutions. The main partners are:

- New York University Libraries
- University of North Texas, The Libraries
- Texas Center for Digital Knowledge

Technical partners will assist in developing an understanding and developing capacity for collecting web-based content and storing, replicating, and importing/exporting the content. Technical partners are:

- San Diego Supercomputer Center at UC San Diego
- Stanford University Computer Science Department
- Sun Microsystems Inc.

The ability to collect and curate web-based content will rely on the knowledge and input from the project's ***curatorial partners***. The curatorial partners include:

- Arizona State Library and Archive
- New York University, Tamiment Library
- University of North Texas, The Libraries
- Stanford University Library's Social Sciences Resource Center
- UC libraries:
    - Institute for Government Studies Library, UC Berkeley

- Institute of Industrial Relations Library, UC Berkeley
- UCLA Online Campaign Literature Archive
- Eight UC libraries that collect CA State and/or Federal publications: UC Berkeley, UC Davis, UC Irvine, UCLA, UC Riverside, UC San Diego, UC Santa Barbara, and UC Santa Cruz.

Project Activities

There are four overlapping paths of activity that run simultaneously throughout the 3-year period of the project.

Content Identification, Selection, and Acquisition (CISA)

There are overlapping paths of activities in the CISA path that will enable curatorial partners to systematically collect and assess web-based collections. The first activity is Content Acquisition and will be led primarily by the CDL. The second activity is Content Identification and Selection and will be led by the UNT. The deliverables from this path are:

- Identification and analysis of web archiving strategies and issues
- Toolkit for curators to assess user needs and plan building of collections
- Framework to conduct and analyze sample crawls
- Case studies, best practices, and collection development guidelines for building web-based collections
- Investigation into extensibility of tools and collection strategies

Content Harvest and Analysis (CHA)

This path builds out and reports on four major application-layer web archiving tools:

- *Curator User Interface (CUI)*: allows authorized selectors to describe and manage web crawls conducted by the *Crawler* tool and to monitor, summarize, verify, and view web crawls using the *Analyzer* tool.
- *Crawler*: harvests web-based content using a primary production crawler, prepares the content for ingest, and hands it off to the storage system. At least one secondary crawler will be used for non-production crawling as needed; for example, to compare with a production crawl in checking for completeness and accuracy.
- *Analyzer*: accesses stored or in-progress web crawls in order to deliver a variety of possible views, including crawl status, error logs, statistical summary pages, and navigable repository-based site mirrors.
- *Export/Import Handler*: allows partners to export (prepare, create, transmit) and import (receive, unpack, and bulk ingest) bulk quantities of curated content (e.g., a collection) in a standardized way.

Content Ingest Retention and Transfer (CIRT)

The Content Ingest, Retention and Transfer (CIRT) path of the project will involve three areas of activity:

- *Data Model*: Specification and implementation of a data model for Web Archive Digital Objects (WADOs) and associated metadata that facilitates their ingest, retention, presentation and interchange
- *Storage*: Evaluation of storage hardware and storage management software for the retention of WADOs; purchase/adoption of selected storage solutions; and development of remote replication strategies
- *Repository*: Modification and enhancement of CDL's Digital Preservation Repository (DPR) to enable the ingest, retention and interchange of objects conforming to the WADO data model;

additional enhancements to the DPR to support the tools being built in the Content Harvest and Analysis (CHA) component of the project.

The CIRT component will be undertaken in five major phases. In the first three phases, the data model, storage and repository aspects of the project are specified, evaluated, developed and integrated with each other and with software deliverables from other components of the project. In the fourth phase, the CIRT component provides the ingest and retention functions for the project-wide collection building effort. In the final phase, remote replication, recovery from remote copies and interchange of WADOs between partner's repositories are piloted.

Partnership Building (PB)

The Partnership Building path will comprise two tasks: 1) The development of essential organizational and operating agreements that contribute to the successful work of the partnership; and, 2) The evaluation of mechanisms that will sustain partner efforts beyond the period of the initial grant.